communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-00775-0

Low responsiveness of machine learning models to critical or deteriorating health conditions

Tanmoy Sarkar Pias[1], Sharmin Afrose[2], Moon Das Tuli[3], Ipsita Hamid Trisha[4,5], Xinwei Deng[6], Charles B. Nemeroff[7] & Danfeng Daphne Yao[1]

Abstract

Background Machine learning (ML) based mortality prediction models can be immensely useful in intensive care units. Such a model should generate warnings to alert physicians when a patient's condition rapidly deteriorates, or their vitals are in highly abnormal ranges. Before clinical deployment, it is important to comprehensively assess a model's ability to recognize critical patient conditions.

Methods We develop multiple medical ML testing approaches, including a gradient ascent method and neural activation map. We systematically assess these machine learning models' ability to respond to serious medical conditions using additional test cases, some of which are time series. Guided by medical doctors, our evaluation involves multiple machine learning models, resampling techniques, and four datasets for two clinical prediction tasks.

Results We identify serious deficiencies in the models' responsiveness, with the models being unable to recognize severely impaired medical conditions or rapidly deteriorating health. For in-hospital mortality prediction, the models tested using our synthesized cases fail to recognize 66% of the injuries. In some instances, the models fail to generate adequate mortality risk scores for all test cases. Our study identifies similar kinds of deficiencies in the responsiveness of 5-year breast and lung cancer prediction models.

Conclusions Using generated test cases, we find that statistical machine-learning models trained solely from patient data are grossly insufficient and have many dangerous blind spots. Most of the ML models tested fail to respond adequately to critically ill patients. How to incorporate medical knowledge into clinical machine learning models is an important future research direction.

Plain language summary

Computational models can be used to evaluate a patient's health condition and predict their risk of dying, for example, in the intensive care unit. These models could be useful to identify patients with worsening health conditions and alert doctors promptly. We test how well several computational models recognize patients with serious or worsening health conditions. We find most of the computational models evaluated cannot recognize critical health events in our tests, which is concerning. Our work highlights the importance of using medical knowledge guided testing to ensure models are suitable, as well as the need to incorporate fundamental medical knowledge into the design of such models.

The Food Drug Administration authorized the first autonomous artificial intelligence (AI) diagnostic system in 2018, which is for detecting diabetic retinopathy[1]. Since then, AI machine learning (ML) based predictive technologies are rapidly made available for incorporation into clinical workflows[2], e.g., for early sepsis detection[3] and predicting surgery time[4]. However, recent studies revealed problems of prediction models under various medical scenarios, e.g., missed detection in mortality prediction or cancer prognosis[5], poor sepsis forecast by a popular U.S. electronic health record software system Epic[6], and models creating incorrect predictive shortcuts for image-based skin cancer detection[7].

These findings point out the urgent need for systematic model evaluation before their clinical adoption to ensure trustworthiness[8]. For example, for in-hospital mortality (IHM) prediction, it is important to measure whether or not ML models can promptly respond to deteriorating patients' conditions. However, due to the immense complexity of the input space, model evaluation is challenging. Exhaustive testing is both unnecessary and impossible in most medical AI applications.

[1]Department of Computer Science and Sanghani Center for AI and Data Analytics, Virginia Tech, Blacksburg, VA, USA. [2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [3]Greenlife Medical College & Hospital, Dhaka, Bangladesh. [4]Banner University Medical Center, Tucson, AZ, USA. [5]University of Arizona College of Medicine, Tucson, AZ, USA. [6]Department of Statistics, Virginia Tech, Blacksburg, VA, USA. [7]Department of Psychiatry and Behavioral Sciences, University of Texas at Austin Dell Medical School, Austin, TX, USA. e-mail: danfeng@vt.edu

The current ML testing practice is very limited in terms of the coverage of disease conditions. Existing model testing is largely restricted to a small percentage (10-15%) of the existing dataset, i.e., test set, as the bulk of the data is reserved for training. Because data imbalance in medicine is common, a typical test set likely has a low coverage of various critical medical conditions and minority prediction class cases. For example, the minority prediction class (i.e., death class) only accounts for 13.5% of an IHM prediction dataset[5]. Even with cross-validation and bootstrapping, the test set is largely limited to the original data.

As a result of the limited test sets, predictive models may be under-evaluated. How they respond to real-world scenarios may be insufficiently assessed. During clinical deployment, new patient conditions could occur out of the distribution of the test set, triggering unexpected failures, e.g., the model failing to produce high enough risk scores for critically ill patients. This issue may disproportionately impact the smaller prediction class, as a typical data-driven model aims to prioritize the accuracy of the majority class samples during training[5]. One approach for increasing test coverage is to use synthetic test samples. Recently, generative technologies have been proposed to produce curated manmade images for testing self-driving vehicles[9,10]. However, image-based solutions do not address the unique temporal challenge in medical time-series applications.

In this work, we develop systematic approaches for generating new test cases beyond the original dataset to assess the responsiveness of ML models to critical health conditions that may occur in clinical settings. Our test case generation is guided by domain knowledge and medical experts. Our experiments involve binary classification tasks, including time-series-based IHM prediction and 5-year breast and lung cancer survivability (LCS) prognosis (Fig. 1). We develop multiple methods for generating high-risk test cases that do not exist in the training data or are underrepresented in the training set. Our solutions can process time series data, which is pervasive in medicine. We also conduct interviews with medical experts to obtain their estimated risks on some of the generated test cases. Our work reveals alarming prediction deficiencies of ML models and points out that ML responsiveness is an important aspect of trustworthiness in digital health.

Methods

Prediction tasks, datasets, and model selection

Our work aims to test medical ML models for their binary classification accuracy under serious disease conditions. We focus on three binary prediction tasks, namely 48-h IHM risk prediction, 5-year breast cancer survivability (BCS) prediction, and 5-year LCS prediction.

The datasets in our study include a 2019 benchmark[11] based on the MIMIC III[12,13] dataset, a 2020 benchmark[14] based on the eICU[15] dataset, and a 2018 reproducibility benchmark[16] based on the Surveillance, Epidemiology, and End Results (SEER) (5-year breast and lung cancer) dataset[16]. The first two datasets contain patients' 48-h time series data in critical care units (ICU). Our study excludes clinical free text notes. As with many medical datasets, the MIMIC-III dataset for IHM, containing 21,139 samples, is imbalanced, with 13.2% death cases (Class 1), and 86.8% non-death cases (Class 0). The eICU IHM benchmark dataset contains a total of 30,681 (88.5% for Class 0 and 11.5% for Class 1) samples with similar attributes and time lengths to the MIMIC III benchmark[14]. Supplementary Fig. S1 shows the distributions of key attributes of both MIMIC III and eICU datasets. The SEER BCS dataset contains 248,751 patient cases with 56 attributes (7 numerical and continuous features and 49 categorical). In the SEER BCS dataset, 12.7% of cases are death cases (Class 0); the rest are survived cases (Class 1). Supplementary Fig. S2 shows the distributions of key attributes. The SEER LCS dataset contains 205,555 cases with 47 features (7 numerical and continuous features and 40 categorical). 84% of patients died in the LCS dataset.

The creation of the MIMIC-III dataset was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Because sensitive health information was de-identified and the dataset did not impact clinical care, the requirement for individual patient consent was waived. The eICU dataset creation is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2). The SEER Program dataset is managed and maintained by the National Cancer Institute (NCI) in the United States. The centralized data collection system enables central IRB submission and approval through reliance agreements with registries. The SEER data collected by registries under state public health reporting authority is HIPAA exempt. MIMIC III is freely available through a proper request to the data source (https://physionet.org/content/mimiciii/1.4/). It requires a license (PhysioNet Credentialed Health Data License 1.5.0), Data Use Agreement (PhysioNet Credentialed Health Data Use Agreement 1.5.0), a training

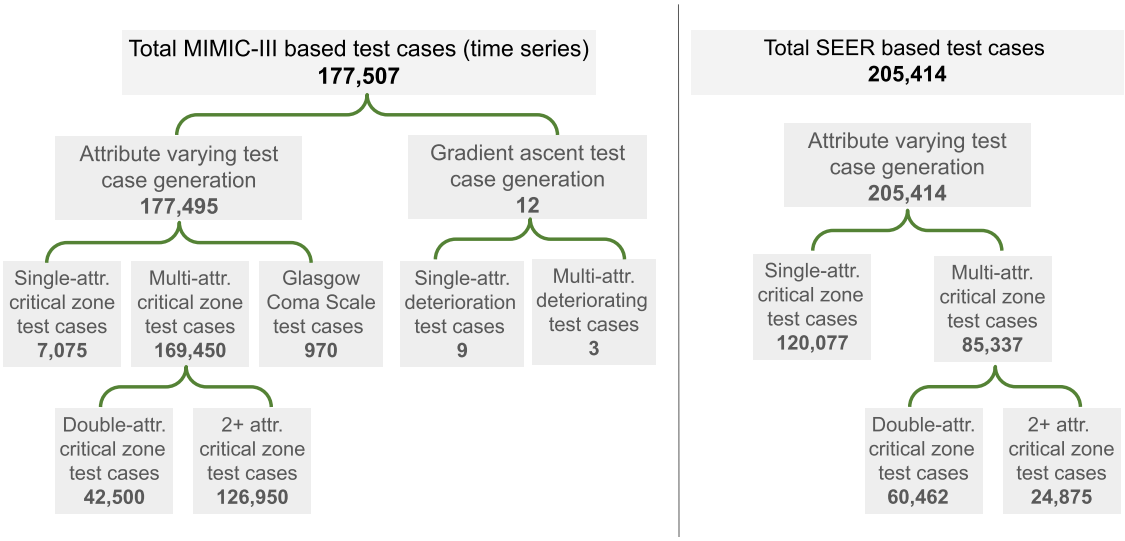

Fig. 1 | Number of generated test cases for evaluating models trained on in-hospital mortality risk prediction and 5-year cancer survivability prediction models. The left side illustrates the generated test case of each category for testing in-hospital mortality risk prediction models trained on MIMIC III or eICU dataset. The right side represents the generated test cases to test 5-year breast cancer survivability (BCS) prediction models. The SEER lung cancer survivability (LCS) models are tested similarly using the single-attribute test cases.

(CITI Data or Specimens Only Research). The eICU dataset can also be accessed (https://physionet.org/content/eicu-crd/2.0/) by completing these mentioned steps. The SEER dataset is also freely available through a proper request to the data source (https://seer.cancer.gov/). It requires the Data Application Form, Data User Agreement, and Acknowledgment of Data Limitations (https://seer.cancer.gov/data/product-comparison.html). The data was accessed through an eRA Commons account, and the data cohort was selected using SEER*Stat software. We gained access to the datasets following the various routes described above. All these datasets are de-identified and public. Thus, an IRB approval is not required for this study, specifically the analysis of de-identified and publicly available data does not constitute human subjects research (U.S. Federal Regulations 45 CFR 46.102).

We select ML models that are commonly used in the medical literature for these prediction tasks. Specifically, we select long short term memory (LSTM) as it is widely used for predicting mortality risk in a 48-h ICU time series dataset—in recent literature[5,17–19]. Similarly, for cancer survivability prediction, we selected multi-layer perceptron (MLP), which was commonly used in analyzing SEER datasets[5,16,20]. In addition, we also evaluated general-purpose ML models commonly seen in medical literature, including XGBoost, AdaBoost, random forest, Gaussian Naive Bayes, and K-nearest Neighbor (KNN). For mortality prediction, we also include channel-wise long short term memory (CW-LSTM) and linear logistic regression (LR) models from the benchmark study[11] and an advanced transformer model.

Dataset preprocessing

We train ML models with benchmark datasets of MIMIC-III[11], eICU[14], and SEER breast and LCS studies[16], following the conventional pre-training processing (e.g., encoding, standardization). As MIMIC-III and eICU benchmark datasets contain missing values, we imputed the values that are missing using the most recent observation (within 48 h) if it exists, otherwise, a value from the normal range of corresponding vitals is mentioned in ref. 11. Masking was used to indicate whether the vital value was original or imputed. The categorical variables, including binary ones, were encoded using a one-hot vector. The numerical features, such as diastolic blood pressure and glucose level, were converted to their standardized form. After preprocessing, each time-series data point became a 76-by-48 matrix (76 computed features and 48 h). The processed dataset was used for training and testing neural network-based models such as LSTM and CW-LSTM models. For non-neural network models that cannot directly process time series, we extracted 6 statistical features (mean, min, max, standard deviation, skew, and number of measurements) from various sub-periods (first/last 10%, 25%, 50%, and full 100%). We did not encode the categorical variables, as they contain values with a meaningful scale. The missing values were replaced with mean values computed on the training set and numerical variables were standardized. In total, we obtained 714 features from each 48-h time series with 17 vitals. The continuous variables were standardized before training. After encoding, the feature vector length of the BCS and LCS datasets became 1418 and 1314, respectively.

Configurations of machine learning models

For IHM risk prediction, we utilized the LSTM model, CW-LSTM model, transformer, LR, AdaBoost, XGBoost, and random forest (RF) models. For 5-year BCS prediction, we used MLP, AdaBoost, XGBoost, and RF models. We utilized the optimal settings of neural network models (i.e., layers, activation, hyperparameters) for each of the tasks from corresponding benchmarks[11,16]. The LSTM model consisted of an input layer (76 dimensions), a masking layer (76 dimensions), a bidirectional LSTM layer (16 dimensions), an LSTM layer (16 dimensions), a dropout layer, and finally a dense layer (1 dimension). In total, the LSTM had 7569 trainable parameters. The CW-LSTM layer consisted of an input layer (76 dimensions), masking layer (76 dimensions), 17 channel layers (for each 17 input features), 17 bidirectional layers (connected to one of the 17 channels layers), another set of 17 bidirectional layers, a concatenation layer connecting all 17 bidirectional layers, bidirectional layer (64 dimensions), LSTM layer (36 dimensions), dropout layer (64 dimensions), and finally a dense layer (1 dimension). In total, the CW-LSTM model had 153,025 parameters. The size of CW-LSTM's parameters was 20 times that of LSTM's. The CW-LSTM model allows independent pre-processing of each variable before combining them. For both LSTM-based models, the optimal hyperparameters are selected using grid search[11]. For example, the batch size, dropout, and time-step are set to 8, 0.3, and 1, respectively. The transformer model consisted of an input layer (76 dimensions), a masking layer (76 dimensions), a positional encoding layer (76 dimensions), 2–3 transformer encoder blocks, a global average pooling layer, a batch normalization layer, a dropout layer (0.3 or 05), a dense layer (32 or 64 units), and finally a dense layer (1 dimension). Each transformer encoder block included a multi-head attention layer with 4 heads (key dimension 76), followed by layer normalization and residual connections. The feed-forward dense layers within each encoder block contained a hidden dimension of 16. In total, the transformer model contains a total of 881,677 parameters (trainable parameters: 293,841, optimizer parameters: 587,684, and non-trainable parameters: 152), larger than the LSTM and CW-LSTM models. For hyperparameter tuning, grid search was employed to select the best hyperparameters (Supplementary Table 1). The LR model was from the sklearn library, utilizing the L2 regularization penalty. To prevent overfitting and to enhance the generalization capability of the model, the parameter C is 0.001. This choice of a small C value effectively controls the amount of regularization applied during training. The remaining hyperparameters were left at their default values, following the standard implementation provided by the Python Sklearn library. This model was trained with the standardized training set. The MLP model used for BCS survivability prediction consists of 2 hidden layers, where each hidden layer contains 20 neurons. The hidden layer used Relu as an activation function. Dropout rate of 0.1 after each hidden layer was used to avoid overfitting. The last layer predicted binary labels using the sigmoid activation function. The MLP model contained 28,831 trainable parameters. MLP hyperparameter is empirically selected using grid-search from a list of predefined values such as the number of hidden layers (1, 2, 3, and 4), number of nodes in each layer (20, 50, 100, and 200), and dropout (0, 0.1, 0.2, 0.3, 0.4, and 0.5)[16]. The other models are implemented using Python's Sklearn library and hyperparameters are tuned using grid search (Supplementary Table 1).

Model training, threshold tuning, and imbalance correction methods

For IHM prediction, LSTM models and transformer models were trained for 100 epochs using the MIMIC-III and eICU datasets separately. For 5-year cancer survivability prediction, MLP models were trained for 25 epochs with the SEER BCS or LCS dataset with optimal hyperparameter settings mentioned[16]. Other models, including XGBoost, AdaBoost, and RF, are trained using the best hyperparameters obtained from grid search (Supplementary Table 1). The models were trained using binary cross-entropy loss. An epoch was selected based on the threshold-agnostic validation area under the precision-recall curve (AUPRC) and validation loss to avoid overfitting. Specifically, we first selected the top 3 epochs with the highest validation AUPRC and then selected the epoch with the minimum validation loss (Supplementary Tables 2 and 3). We monitored the validation loss and training loss difference to prevent overfitting. In all experiments, the chosen ML model demonstrated a small loss difference (Supplementary Tables 2 and 3).

Besides evaluating models trained on the original training sets, we also experimented with resampling and reweighting techniques and measured how well the resulting bias-corrected ML models performed in our critical zone tests. The reweighting technique has demonstrated superior performance in healthcare datasets, as evidenced by prior studies[21]. For resampling, we tested two generative resampling approaches, SMOTE (Synthetic Minority Oversampling Technique) and AdaSyn (Adaptive Synthetic Sampling). We employed Python's Imblearn library to apply SMOTE and AdaSyn oversampling techniques, generating balanced training sets by increasing samples from the minority class (sizes shown in Supplementary

Table 4). For reweighting, we utilized Python's Sklearn library to compute balanced class weights based on the training sets (Supplementary Table 5). These methods are applied to the LSTM model for mortality prediction and to the MLP model for cancer survivability prediction.

The training, validation, and test set breakdown for MIMIC-III and eICU datasets is 70%, 15%, and 15% and 80%, 10%, and 10% for the BCS and LCS datasets. After model calibration, a threshold-tuning process is conducted on the validation set, and an optimal threshold is selected based on balanced accuracy and F1 score for the minority class. Specifically, after training, we first conducted model calibration by applying Isotonic Regression using the validation set. Model calibration mapped the predicted probabilities to actual probabilities. Then, we performed threshold tuning to determine the optimal threshold. The minority F1 score and balanced accuracy were computed on the validation set for each threshold ranging from 0.0 to 1.0 with a step size of 0.01. Subsequently, the top three thresholds yielding the highest minority F1 scores were identified, and the optimal threshold maximizing balanced accuracy across all validation samples was selected. This process was repeated for 3 independently trained models of each type, and the average threshold was calculated from these independent trials. Thresholds are shown in Supplementary Table 6. The tasks were executed on a machine with Ubuntu 18.04 operating system, x86-64 core-i9 architecture, 8 physical cores (16 virtual cores), and 32 GB RAM. The experimental code and models were written using Python 3.7, TensorFlow 1.15, and Keras 2.1.2. The cancer survivability prediction MLP model was trained on a machine with x86_64 Intel(R) Xeon(R) CPU 2.40 GHz (40 cores) and 125 GB RAM. The experimental code and model were written using Python 3.6, TensorFlow 2.9.0, and Keras 2.9.0.

Mapping neuron activations

We visualized the activated neurons in a neural network model for a particular input. The Keras backend was used to capture the neuron outputs from the bidirectional layer output and LSTM layer output for the mortality risk prediction model. Sigmoid activation was applied to obtain neuron output values in the range of [0, 1]. To quantify changes in neuron activation, we defined and computed Neural Zone Activation (NZA) and average zone difference ΔNAZ. A zone is defined by the attribute range bounded by two values. NZA calculates the average neurons' activations within a zone, where a zone can be a critically low, critically high, or normal range (Supplementary Equation 1). ΔNAZ computes the average NZA difference between two zones (Supplementary Equation 2), such as normal and critically high zones, indicating how much neurons react to zone changes. There is no standard value for ΔNAZ. A relatively higher value indicates a good response.

Statistical methods

Model performance is reported using the average and standard deviations, which are calculated using 9 or 15 trials. The trials were performed using 3 model instances that have identical architecture and were trained on the same training set with random model parameter initialization. Each of the 3 model instances is evaluated with 3–5 test sets. The distribution shift of the synthesized test dataset from the original training sets was quantified by Wasserstein distance (WD)[22,23]. We used an implementation from the Python library called scipy.stats.wasserstein_distance. First, the WD was calculated between the same features from the whole original dataset and the synthesized test set. Then, the feature-specific WD was averaged to obtain the mean WD for quantifying the distribution shift.

Attribute-based test case generation for in-hospital mortality risk prediction

We created new cases by increasing or decreasing one or multiple vital health parameters in the seeding records. To reduce computing complexity, we prioritized by focusing on the most influential features. Relevant medical terminologies are explained in the Supplementary Notes.

In the single-attribute variation, *we* generated new test cases by varying a single attribute at a time while keeping other attributes unchanged. We then evaluated how the model reacts to these changes and its ability to recognize associated risks (e.g., hypoglycemia). Specifically, given an attribute *A*, single-attribute variation for time series involved the following operations. First, we identified *A*'s minimum and maximum values in the MIMIC-III or eICU datasets, which defined the observed range. Then, the mean and the variance of attribute *A* were computed from the entire dataset. Using the variance and the observed range, we generated a series of random values for every value from that range, one value for each of the 48 h. Then, the new test case was formed by having these generated values for attribute *A* and other attribute values directly inherited from the seed. We repeat this process for every possible attribute value from the observed range with step 1.

Multi-attribute variation generated new test cases by modifying two or more attributes, aiming to represent medical conditions that were characterized by variations in multiple related attributes. We further differentiated two scenarios: (a) a single set of medically correlated attributes driven by one underlying disease condition, e.g., high diastolic and systolic blood pressure due to hypertension, and (b) medically correlated attributes due to multiple underlying conditions, e.g., hypertension and diabetes. These test cases were used to assess the ML model's ability to respond to the risks of multiple disease conditions in patients. One of the test sets was created by changing multiple vitals such as systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature at the same time. A test case was assigned a ground truth label using existing literature or under the guidance of medical doctors. 6 multi-attribute test cases and 12 deteriorating test cases were directly labeled by the medical doctor (Supplementary Tables 7 and 8).

Deteriorating test case generation for MIMIC-III

We leveraged the gradients of LSTM to guide the generation of new test cases. This method is automatic and does not require the specification of attributes to change, aiming to generate new test cases that are challenging for ML models to classify correctly. Such cases typically occur at the decision boundary of the classifier. Our method started from a healthy patient's record (i.e., a seed with low or zero mortality risk). The seed is a time-series record far away from the classifier's decision boundary. We incrementally adjusted the attribute values of the seed by following the steepest direction (i.e., gradient) that can maximize the loss (i.e., prediction errors of the ML model). This process explores the local hyperspace and iteratively produces new cases that are closer and closer to the ML model's decision boundary. Computationally, given a trained ML model and a healthy patient's time series record as the seed, we computed the derivative of the model's loss function, i.e., gradient (Supplementary Equation 3). The gradient is a vector of partial derivatives describing the direction and rate of changes of the loss function. Then, we changed the test case in the direction of increasing gradient. Our algorithm is described in the Supplementary Methods Section. The step size or learning rate to control the magnitude of the change was set to 0.001–0.2 in our experiment depending on the attribute (Supplementary Table 9). Our method focuses on generating samples; it differs from the common gradient descent process, which adjusts model weights to minimize loss. We have two ways of creating gradient-based test cases from the MIMIC-III dataset—single-attribute gradient approach and multi-attribute gradient approach. In the former, we focus on a single attribute and apply gradient ascent solely to modify that specific attribute. This approach allows one to observe the individual impact of each attribute on the mortality risk. In the latter, we simultaneously change values of multiple attributes using gradient ascent. Gradient approaches create test cases that represent deteriorating health conditions in continuous time series. Supplementary Table 10 shows the various categories of test sets and their sizes.

Glasgow coma scale test case generation

The Glasgow Coma Scale (GCS)[24] is a neurological scale that assesses a patient's level of consciousness. It evaluates responses in three categories: eye-opening (E), verbal response (V), and motor response (M), adding up to a score ranging from 3 to 15. A lower score indicates a more severe

impairment of consciousness. Definitions of values in each category are given in Supplementary Table 11. A GCS score can be representative of multiple sets. For example, GCS total 10 can be the outcome of (E, V, M) = (3, 3, 4), or (4, 4, 2), etc. The GCS total test set contains all the possible combinations of (E, V, M) for each particular GCS total value. The double attribute-based GCS cases were also created by varying both attributes and keeping the other constants to healthy values.

Attribute-based test case generation for 5-year cancer prognosis

Single-attribute variation. Similarly, we engineered cancer test cases by varying one attribute of a seed record. The attribute may be the size of the tumor (T), the number of positive lymph nodes (N), the number of examined lymph nodes (ELNs), or the grade of the cancer cell. T and N are the two most important factors for determining cancer severity or stage[25]. T has 4 categories based on the size. The tumor test set was created by varying the size of 3 seeds in the surviving class, using the range (0–986 mm) from the original SEER dataset. This BCS tumor size test set contains 12,891 cases, including 18 T0 cases, 243 T1 cases, 390 T2 cases, and the rest of 12,240 T3 cases. The LCS tumor size contains 8367 cases, including 12 T0 cases, 171 T1 cases, 273 T2 cases, and 7911 T3 cases. (T4 cases cannot be created, as it is not associated with a quantitative value). The number of positive lymph nodes (N) is divided into 4 categories. The positive lymph node test case was created similarly by changing the corresponding value from the same 3 seeds using the attribute range (0–84). For BCS, we generated 7686 test cases, including 90 N0 cases, 270 N1 cases, 546 N2 cases, and 6780 N3 cases. For LCS, we generated 24,264 test cases, including 333 N0 cases, 999 N1 cases, 1998 N2 cases, and 20,934 N3 cases. Supplementary Notes have more details of T and N category definitions.

The ELN test case was created similarly by varying the number of ELNs (range in [0, 86]) from 3 seeds and keeping other values the same as the seeds. The ELN test set contains a total of 3510 cases for BCS and 1835 cases for LCS. Although the number of ELNs is not directly related to the cancer staging, it is crucial for diagnosing cancer. Several studies proposed that there should be a standard (or a minimum) number of ELN cancer diagnoses[26–29]. The grade of the cancer cell represents the spreading and growth intensity of the cancer cell[25]. The SEER dataset contains 1–4 grades where the higher grade represents faster growth and speed and another grade 9 for undetermined (not stated/applicable). For BCS, we created test sets for each of 1–4 grades, where each set contains 24,875, created from 21,723 cases from the majority Class 1 (survival) and 3152 cases from the minority Class 0 (death). We utilized the entire validation set as the seed pool, allowing for a more comprehensive evaluation. In total, the 1–4 grade test set contains 99,500 cases (24,875 cases for each grade). To create each grade test set, we set the corresponding grade value to all data points in the validation set.

Double- and triple-attribute variations. Double-attribute variation generated new BCS test cases by changing a pair of attributes from the 3 continuous attributes, which are the size of the tumor (T), the number of positive lymph nodes (N), and the number of ELNs. The grade attribute was excluded, as it is categorical. The tumor size and positive lymph node combination test set contains 18,531 test cases. The tumor size and number of ELNs combination test set also contains 18,531 test cases. The number of ELN and positive lymph node combination test sets contain 23,400 test cases. The triple-attribute test set was created by setting three attributes simultaneously to represent serious disease conditions, e.g., tumor size to T4, number of positive lymph nodes to N3, and grade to 4. The validation set, consisting of 24,875 cases including 21,723 cases from Class 1 (survived) and 3152 from Class 0 (death), was used as seeds. While the tumor size and number of positive lymph nodes are continuous variables, we treated them as categorical by selecting a value from the T4 and N3 range respectively. As a result, the triple-attribute test set contains 21,723 cases derived from Class 1 seeds and 3152 cases derived from Class 0 seeds. Supplementary Table 12 summarizes the various categories of test sets and their sizes. We performed double- and triple-attribute variation tests for BCS models, not on LCS models.

For labeling generated breast and lung cancer test cases, we used authoritative literature to assign labels. We labeled cases with Class 0 (indicating low survivability) if there was a strong presence of cancer (i.e., T 1–3, N 1–3, and grade 2–4). For ELNs, the previous studies using SEER datasets[30,31] suggested using at least 8-9 ELNs for stage T1 diagnosis, 37 ELNs for T2 diagnosis, and 87 ELNs for T3 diagnosis. As ELN is not directly responsible for the death, that attribute was not considered during labeling.

Selection of seeds

We used existing patient records from the original dataset as seeds (i.e., starting points) to generate synthetic test sets. We selected seeds from the IHM dataset that are real-world non-death patient cases that exhibit healthy attribute values. Seeds were chosen as follows. For attribute-based test case generation, we randomly selected seeds from MIMIC-III Class 0 (survival case) following two criteria. First, the mean (of 48 h) attribute values are within the range of ideal health conditions defined in Supplementary Table 13. In addition, the standard deviation of each attribute needs to be less than or equal to the mean standard deviation (Supplementary Table 14) of the MIMIC-III dataset. Our evaluation of attribute-based test case generation involved 5 seeds and the statistics of these 5 seeds are given in Supplementary Table 15. The deterioration test case generation involved another 3 seeds, which were selected randomly from Class 0 of MIMIC-III. Since the eICU dataset contains similar samples with identical features and a consistent 48-h time duration, we utilized the same test set generated from MIMIC to evaluate models trained on the eICU dataset. Additionally, the selected seed attributes fall within the healthy (ideal) range, minimizing the out-of-distribution effects on models trained on the eICU dataset. For the cancer survivability prediction task, test cases involving changing a numerical variable were generated from 3 randomly selected seeds from the surviving class. Test sets are separately generated from each of the SEER BCS and LCS datasets. Test sets involving categorical variables, such as grades test and triple-attribute test sets, were generated using all validation data points from SEER.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

For IHM prediction, we generated 177,507 new time-series test cases based on MIMIC-III to represent serious patient conditions and used them to evaluate the responsiveness of machine-learning models (Supplementary Table 10). 126,950 cases are generated by modifying multiple vital attribute values in 5 seed records, 42,500 cases by modifying double attributes in seed record, 7075 cases by modifying a single attribute value in a seed record, 970 cases by modifying GCS, and 12 cases by gradient ascent. Modifications to vital attributes are bounded by the minimum and maximum values of the attribute in the IHM datasets and focus on critically high and critically low ranges of the 6 vitals. We carefully use literature[32–36] to identify these ranges (Supplementary Table 13). The test case generation also ensures the continuity of the time series. The 6 types of attributes include systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature. A seed record is a real-world patient case selected from the MIMIC-III dataset that is a non-death case whose attributes are in the typical healthy ranges[33–36]. The other 12 test cases are generated using a gradient-ascent approach, which modifies the seed by following the direction of the steepest increasing loss function.

Each synthetic test case is assigned a label, death (Class 1) or survival (Class 0) for IHM prediction. Labels are verified either by a medical doctor or confirmed by the literature. These labels are considered ground truth in our study. Two medical doctors reviewed 18 generated test cases (6 attribute-based cases and 12 gradient-based cases), where the test cases are time series data and the risk scores of the medical doctors' output are quantitative, between 0 and 1. The medical experts estimated risk values are in Supplementary Tables 7 and 8. Labels of the other 177,489 test cases are

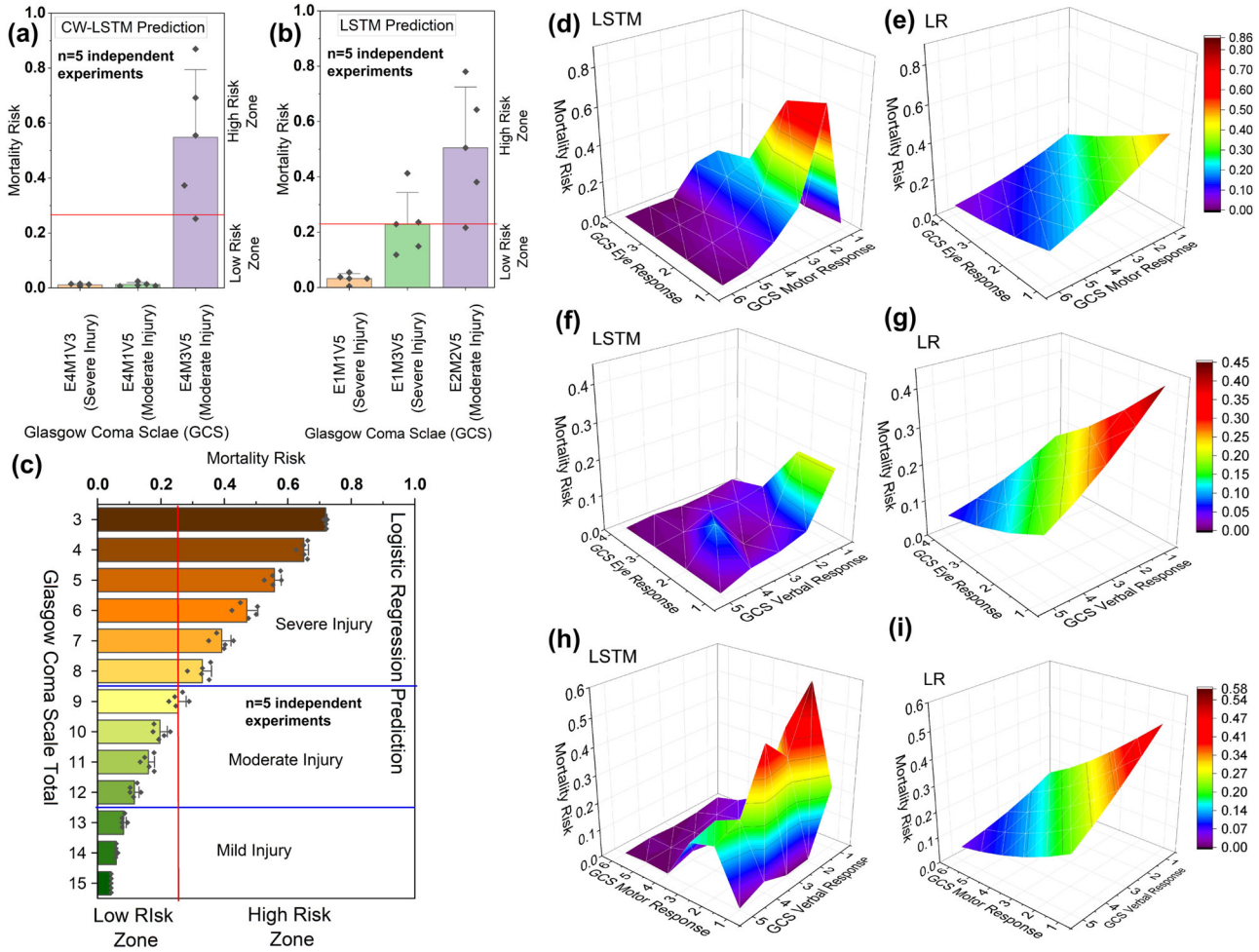

Fig. 2 | Mortality risk (MR) prediction for Glasgow Coma Scale for different combinations using three machine learning models. MR predicted by (**a**) channel-wise LSTM model for three injury cases (E4M1V3, E4M1V5, and E4M3V5) represented by 3 bars and the X-axis, **b** LSTM model for three injury cases (E1M1V5, E1M3V5, and E2M2V5) represented by 3 bars and the X-axis, and **c** Logistic regression for injury cases for all combinations of GCS scores (GCS total 3–15) represented by the bars and the Y-axis. The error bar represents standard deviations calculated from five independent experimental trials (models). MR prediction of injury cases defined by different combinations of GCS eye and motor response scores by **d** LSTM and **e** logistic regression model. MR predicted by **f** LSTM and **g** logistic regression using injury cases defined by different combinations of GCS eye and verbal response scores. MR prediction of injury cases defined by different combinations of GCS motor and motor response scores by **h** LSTM and **i** logistic regression.

inferred based on expected ranges of vital health parameters of healthy individuals extracted from medical literature. Synthetic test cases persistently containing vital values in critical zones represent patients in sustained critical health conditions, and thus are labeled Class 1. These cases should receive a high mortality risk prediction from ML models. We define ML responsiveness as the model's ability to react to substantial changes in input values, e.g., by increasing the mortality risk score for IHM prediction.

Based on the SEER 5-year BCS dataset, we generated 205,414 test cases to represent different patient conditions. Among them, 120,077 cases are generated by changing single attributes (including 7686 cases representing different N stages, 12,891 cases for the T stage, and 99,500 cases for grades), 60,462 cases by modifying double attributes, and 24,875 cases by changing triple attributes from the seed cases. Based on the LCS dataset, we generated three sets of single-attribute test cases totaling 31,136 cases, which include 8367 cases representing different T stages, 24,264 cases representing N stages, and 1835 cases representing ELNs (Supplementary Table 12). We manually assigned labels to synthesized test cases guided by the literature[25–31].

ML performance under Glasgow Coma Scale (GCS) testing

For IHM prediction, we assess MIMIC III-based LSTM, CW-LSTM, and LR models with test cases containing varying GCS scores (Fig. 2), including severe injury cases with GCS scores 3–8, moderate injury with 9–12, and mild or no injury with 13–15. A low GCS score indicates a poor health condition[24] (medical meanings of each category are shown in Supplementary Table 11). The CW-LSTM model gives near zero mortality risk values for 15 severe injury cases, for example, E4M1V3 in Fig. 2a, i.e., a case with an eye response score of 4 out of 4, a motor response of 1 out of 6, and a verbal response score of 3 out of 5. For a moderate injury case E4M1V5, CW-LSTM also gives an unexpectedly low mortality risk (0.01) prediction, i.e., predicting the healthy outcome of the patient. The model's prediction is inconsistent, as another moderate injury case E4M3V5 receives a high mortality risk of 0.58.

Similar inaccuracies and inconsistencies are also observed for the LSTM (MIMIC III) model tested. For instance, the LSTM model mistakenly considers a severe injury case E1M1V5 to be much more likely to survive than a moderate injury case E2M2V5 (Fig. 2b). In contrast, the LR model consistently predicts at least 0.3 mortality risk for severe injury cases and responds well (Fig. 2c). For mild injury cases, the LR model consistently predicts a low mortality risk. The 3D surfaces of the LR model appear smooth and the model reacts to decreased eye and motor signals (Fig. 2e). In contrast, LSTM's 3D plots are less monotonic, exhibiting bumps (Fig. 2). For the most severe cases (subscores being 1 or 2), LSTM's risk predictions incorrectly drop.

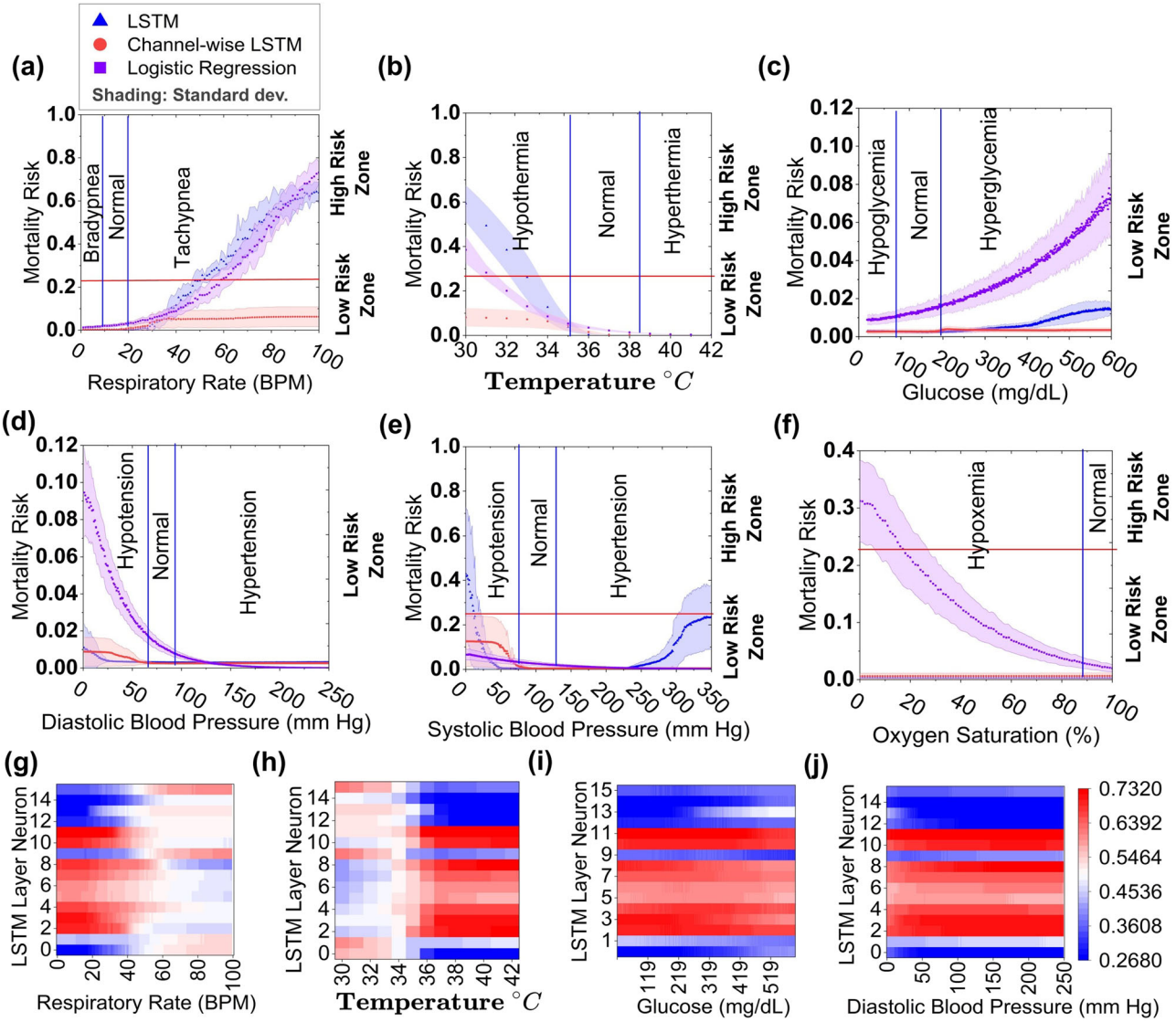

Fig. 3 | Mortality risk prediction for single vital-sign tests using three machine learning models (LSTM, Channel-wise LSTM, and Logistic Regression) and visualizing the neural activation map of the LSTM layer consisting of 16 neurons. LSTM, Channel-wise LSTM, and Logistic Regression (LR) predict the mortality risk (MR) of **a** respiratory rate, **b** body temperature, **c** glucose, **d** diastolic blood pressure, **e** systolic blood pressure, and **f** oxygen saturation test sets (synthesized). The mortality risk (MR) is represented by X-axis and MR above and below a red horizontal line (threshold = 0.22) indicates a high or low mortality risk zone, respectively. The shading represents the standard deviation calculated from five independent experimental trials (models). The entire range of each vital sign (except oxygen saturation) value is divided into three segments, low, normal, and high, by the blue vertical lines. The low and high values within these ranges indicate critical health conditions. Figures in the last row (**g**–**j**) represent the neural activation map. These are the neural activation values, calculated after applying the sigmoid function, when the model is fed with test cases varying a single vital, such as **g** glucose, **h** diastolic blood pressure, **i** temperature, and **j** respiratory rate.

ML performance under critical zone tests

Single-attribute critical zone test results. We evaluate the MIMIC III-based LSTM, CW-LSTM, and LR models' ability to respond to a single deteriorating attribute while keeping other attributes stable as in the seed (Fig. 3). The CW-LSTM model fails to recognize bradypnea, i.e., an abnormally slow breathing rate, and gives only slightly elevated mortality risk prediction (mean mortality risk 0.05 and standard deviation 0.04) for tachypnea, i.e., rapid breathing (Fig. 3a), insufficient to trigger an alert. Similarly, CW-LSTM is unable to recognize most of the abnormal vitals. Its mortality risk prediction gives a negligible change to an abnormal patient's glucose level (Fig. 3c) and oxygen saturation rate (Fig. 3f). For the other 3 attributes tested, CW-LSTM gives small partial responses to either a high critical zone or a low critical zone, but not both. For example, it is unable to recognize high body temperature anomalies and only slightly raises the mortality risk to 0.01 to 0.08 (standard deviation 0.033) for severe hypothermia below 34 °C (Fig. 3b), which is still much below the classification threshold (0.22). CW-LSTM's response to abnormal diastolic and systolic blood pressure is also inadequate (Fig. 3d and e).

The LSTM model gives much more elevated risk prediction than CW-LSTM for tachypnea (Fig. 3a) and hypothermia conditions (Fig. 3b). Out of the 3 models tested, LSTM is the only machine-learning model that responds to both systolic hypotension and hypertension conditions, producing a U-shaped curve (Fig. 3e). However, LSTM consistently gives an ultra-low risk prediction for abnormal diastolic blood pressure (Fig. 3d). Similar to CW-LSTM, LSTM does not recognize hypoxemia, i.e., low blood oxygen level (Fig. 3f), bradypnea, hyperthermia, and abnormal glucose level (Fig. 3c), exhibiting either monotonic or near-flat risk prediction curves insensitive to abnormal vitals. Compared to the other models, LR gives a substantially higher risk prediction for hypoxemia. It also computes elevated risk scores in response to increasing hyperglycemia and diastolic hypotension conditions. For all attributes, LR is only able to recognize one end of the critical zones, but not both. Overall, LR, LSTM, and CW-LSTM correctly

predict 37.7%, 37.8%, and 22.4% of the single-attribute critical zone test cases on average (Supplementary Table 16).

Neuron activation analysis

We visualized neuron output from intermediate layers of the MIMIC III-based LSTM model. Neurons whose activations change with changing variable values are the responsible neurons for recognizing that variable. We found most of the LSTM neurons have low or no responses to varying glucose and diastolic blood pressure values (Fig. 3i and j). In contrast, neurons are more responsive to temperature and respiratory rate changes, e.g., sharp changes in all neuron activation between 34 °C to 36 °C (Fig. 3h) and around or above 40 bpm (Fig. 3g). However, neurons exhibit minimal or no changes in activation for higher temperatures or for critically low respiration rates.

To quantify changes in neuron activation, we computed NZA and average zone difference ΔNAZ, new metrics defined by us (Supplementary Equations 1 and 2). NZA averages neurons' activations within a zone, where a zone is critically low, critically high, or normal range (Supplementary Equation 1). ΔNAZ computes the averaged NZA difference between zones (Supplementary Equation 2), indicating how much neurons react to zone changes. The LSTM model shows low (0.01–0.04) ΔNAZ in most cases (Supplementary Table 17). In a few cases, e.g., temperature ΔNAZ (low, normal) and respiratory rate ΔNAZ (high, normal), the values are relatively high (0.14–0.16).

Multi-attribute critical zone test results

We evaluated the 3 MIMIC III-based ML models under 42,500 double attribute varying test cases (Fig. 4), including respiratory rate and heart rate pair (first row), systolic and diastolic blood pressure pair (middle row), and glucose and diastolic blood pressure pair (third row). The CW-LSTM model does not generate high mortality risk predictions for most critical zone cases, consistent with its single-attribute test performance in Fig. 3. The LR model gives better performance than CW-LSTM, predicting higher risks for some critical zone combinations (Fig. 4a, d, g). However, its prediction is monotonic, thus, unable to recognize both high and low critical zones of an attribute pair. For example, LR fails to alert when patients have low respiratory rate and low heart rate. LSTM model exhibits prediction behaviors (Fig. 4b, e, h) consistent with its single attribute performances in Fig. 3.

In a 6-attribute varying test setting, we evaluated the responsiveness of MIMIC III-based ML models under 6 changing vitals, where test cases have abnormal systolic and diastolic blood pressures, blood glucose level, respiratory rate, heart rate, and body temperature values in their respective critical zones. We recorded how much mortality risk scores changed and showed the distributions in Fig. 4j–o. Figure 4j–k show the mortality risk difference (MR) between each high critical zone test case and its corresponding seed. Medically speaking, the risk should increase under worse health conditions. The CW-LSTM model consistently predicts high mortality risk for most cases, resulting in a positive MR for over 90% of the cases (Fig. 4i, o). The LR model (Fig. 4j) produces negativeMR for all cases, which is incorrect. The LSTM model generates positive MR for two-thirds of the 12,694 test cases.

The 3 models under low critical zone tests performed similarly (Fig. 4m–o), where CW-LSTM responds to multi-attribute critical conditions the most effectively and LR the least. Overall, LR, LSTM, and CW-LSTM correctly predict 6.2%, 45.7%, and 69.3% of the multi-attribute critical zone test cases on average (Supplementary Table 16), respectively.

Results on test cases with deteriorating conditions

For IHM prediction (using the MIMIC III dataset), we used a gradient ascent method to generate 12 time-series test cases with deteriorating health conditions. 9 of the 12 test cases contain one vital that worsens during the 48 h and is in the critical zone during the last 24 to 48 h, including 3 cases of decreasing systolic blood pressure (Fig. 5a), 3 cases of increasing respiratory rate cases (Fig. 5b), and 3 cases of decreasing body temperature (Fig. 5c). The other 3 test cases have multiple (3) worsening vital signs, with vitals being in critical zones during the last hours (ranging from 30 to 48 h). All 12 test cases should receive a high mortality risk prediction, i.e., Class 1. We confirmed these labels with two medical doctors who manually reviewed the time series data.

Average mortality risks predicted by ML models on the 9 single-attribute deteriorating test cases are in Fig. 5d. Out of the 9 single-attribute deteriorating test cases, LR only detects 2 (22%) respiratory rate cases (average risk 0.38) and fails to detect the other 7. CW-LSTM (MIMIC III) detects 4 (44%) out of the 9 deteriorating test cases, including 2 systolic BP cases (average risk 0.32) and 2 respiratory rate cases (average risk 0.39). However, it is unable to detect deteriorating temperature cases. LSTM reports 5 (56%) out of the 9 deteriorating test cases, including 2 systolic BP cases (average risk 0.22), 2 temperature cases (average risk 0.23), and 1 respiratory rate case (risk 0.22). Collectively, the models detect 41% out of single-attribute deteriorating test cases.

Multi-attribute test cases have 3 deteriorating vitals, including oxygen saturation, temperature, and diastolic blood pressure. The LSTM (MIMIC III) model detects all 3 cases (average risk 0.23), whereas CW-LSTM fails to generate any alerts. LR (MIMIC III) issues alert for 2 out of 3 cases (average risk 0.56). Collectively, the 3 models detect 56% out of the multi-attribute deteriorating test cases. Overall, the models' average accuracy under all deteriorating test cases is 44%.

5-year cancer survivability results

We found similar deficiencies in the ML model, in terms of the model's ability to respond to test cases representing serious cancer conditions.

Single-attribute test results

We evaluated the responsiveness of MLP models trained on the BCS dataset to a single deteriorating attribute (Fig. 6 and Supplementary Table 18) while keeping other attributes the same as the seed. The BCS-MLP model shows some responsiveness with varying tumor size (Fig. 6a), however, remains above the survivability threshold (0.71) in all cases. As a result, the model fails to trigger an alert for tumor sizes representing critical stage T1 (tumor size less than 20 mm) to T3 (tumor size larger than 50 mm). On the other hand, the model triggers alerts for 74.4% of the 6,780 N3 stage (Fig. 6b). It fails to generate any alert for 270 N1 and 546 N2 stage cases. The BCS-MLP model accurately generates alerts for 66.4% of critical cases (N1-N3). It decreases the survivability for lower numbers of ELNs, however, still fails to trigger any alerts (Fig. 6c). Each of the grade test sets (1–4) is generated from 21,723 surviving patient seeds from the original dataset and the results are in Fig. 6g. Out of these test cases, the BCS-MLP model generates insufficient numbers of alerts, only for 4.6% of the grade 2 cases (slower-growing cancer and less likely to spread), 7.4% of the grade 3 cases (growing cancer), and 6.7% of the grade 4 cases (faster-growing cancer and likely to spread). For tests generated from 3152 death events, the model generates more alerts, 57.9% of the grade 2 cases, 65.8% of the grade 3 cases, and 63.9% of the grade 4 cases. The model did not generate any alerts for 97% and 48.7% of grade 1 test cases generated from seeds of survived and death events, respectively. The LCS-MLP model shows higher responsiveness with variations in tumor size (Fig. 7j), generating alerts in 80.1% of the test cases (Supplementary Fig. S3). It also responds well to varying positive lymph node numbers (Fig. 7k) with alerts generated in 92.9% of the cases. However, BCS and LCS models do not react to the increasing number of lymph nodes examined (Fig. 7i, l).

Tree-based ensemble methods, including AdaBoost, Random Forest, and XGBoost, produced somewhat similar results across all four datasets (Supplementary Fig. S4). The RF model demonstrates good responsiveness for most attributes in both MIMIC III and eICU datasets. However, it does not respond to critically high respiratory rates. XGBoost and AdaBoost have little reaction to attribute changes. In cancer survivability tasks, none of the models responds to worsening patient conditions, except AdaBoost for LCS-positive lymph nodes.

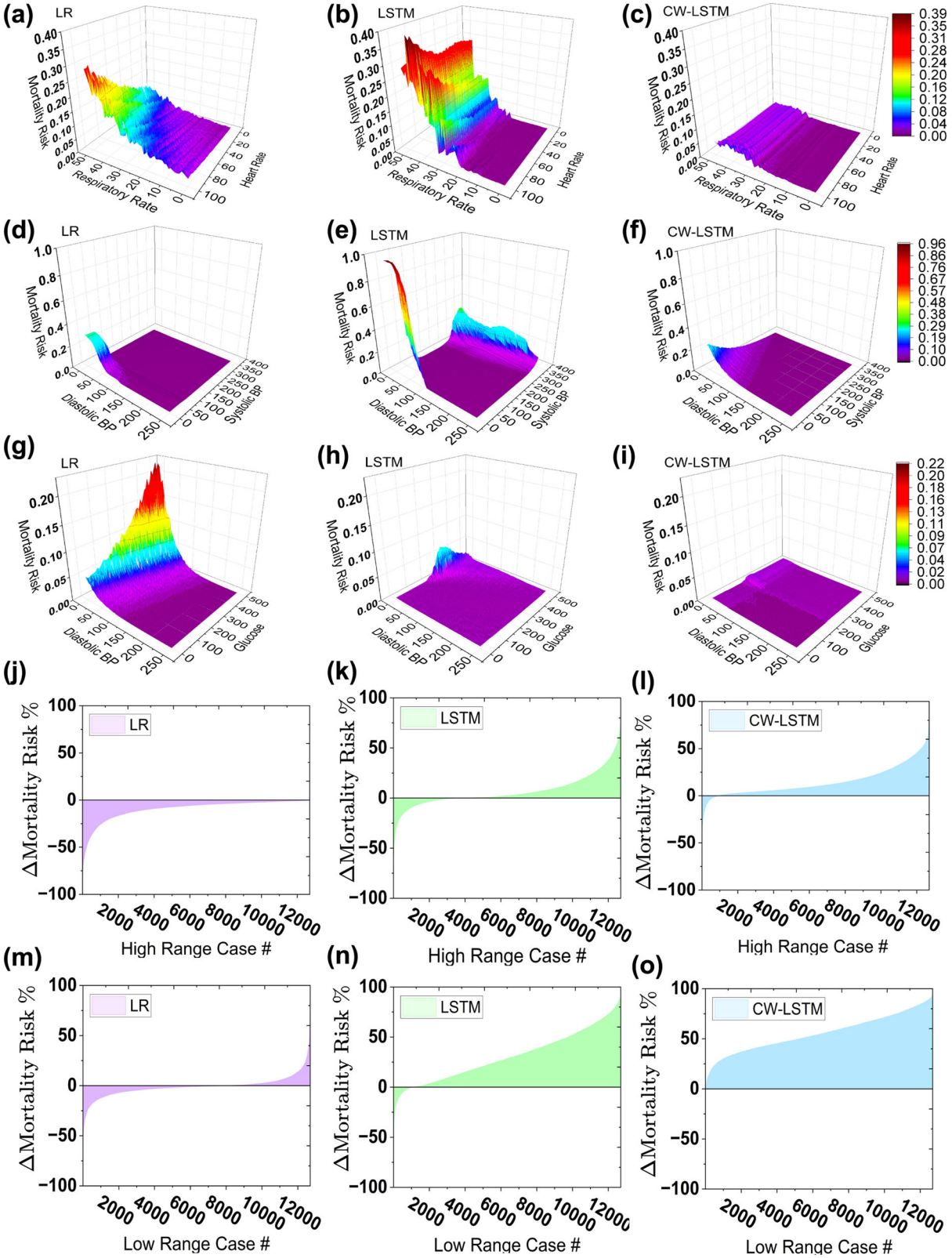

Fig. 4 | Multi-attribute test results, including the mortality risk prediction under double-attribute variation tests and the mortality risk difference (ΔMR) between the seed and 6-attribute variation test cases. Risk prediction under varying respiratory rate and heart rate by **a** logistic regression, **b** LSTM model, and **c** CW-LSTM model. Risk prediction under varying systolic and diastolic blood pressure by **d** logistic regression, **e** LSTM model, and **f** CW-LSTM model. Risk prediction under varying glucose and diastolic blood pressure by **g** logistic regression, **h** LSTM model, and **i** CW-LSTM model. **j**–**l** represent ΔMR for high critical range cases and **m**–**o** represent ΔMR for low critical range cases. The test set is generated by simultaneously varying systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature and values are randomly selected from the critical zone. The graph shows the mortality risk difference (ΔMR) calculated by subtracting the predicted mortality risk of the seed (Class 0) from the predicted mortality risk of its corresponding critical case. The X-axis represents the case numbers and the Y-axis represents ΔMR. It is expected to get a positive MR difference and the negative ΔMR cases represent failed test cases.

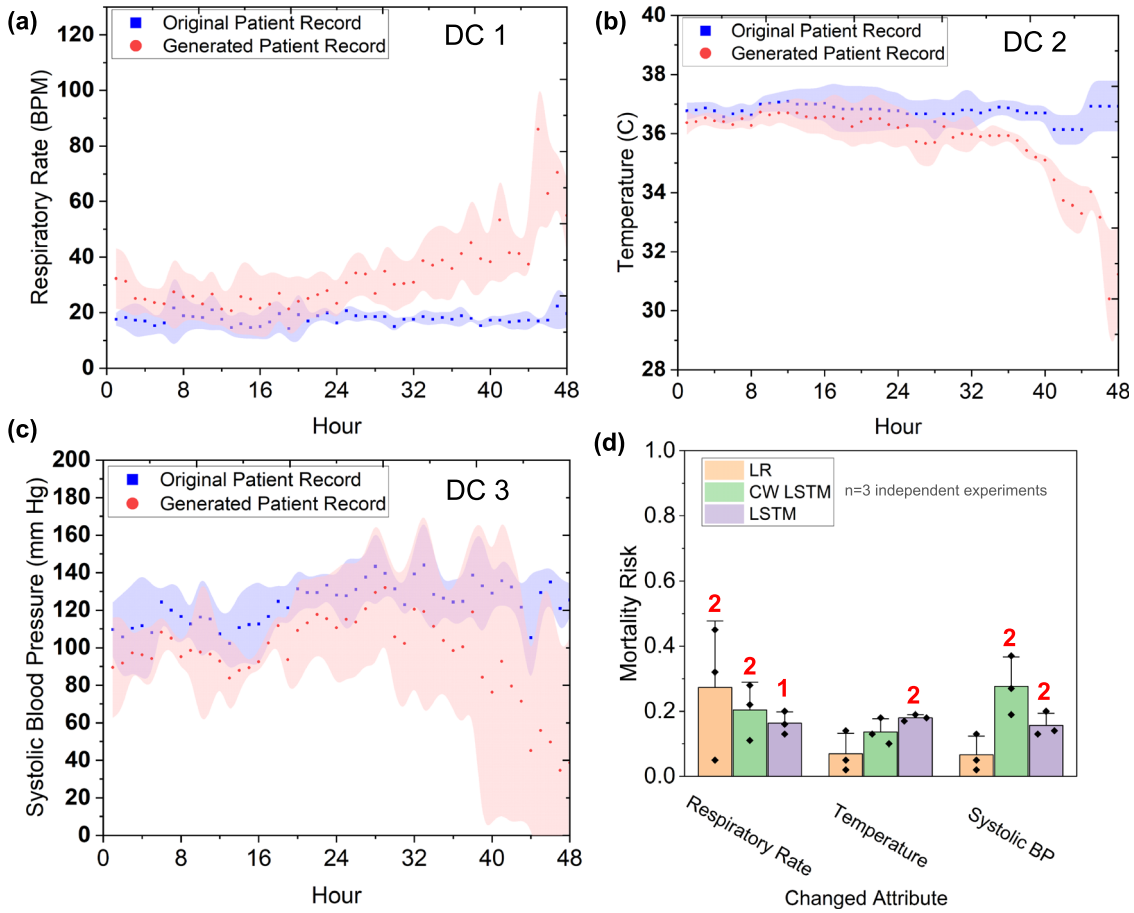

Fig. 5 | Gradient-generated deteriorating test cases and machine learning models' mortality risk predictions by LR, CW-LSTM, and LSTM models. a–c show the average time series of the generated abnormal test cases (in red area curves) and the normal seed cases used (in blue area curves) for each of the 3 attributes. **d** Models' predicted average mortality risks for each deteriorating attribute. The standard deviation is indicated by the error bar and the number of detected cases out of 3 is shown in red.

Double-attribute test results

The BCS-MLP model was tested under 60,462 double attribute varying test cases, including (a) tumor size (T) and positive lymph node (N) combination test (Fig. 6d), (b) number of ENL and positive lymph node (N) combination test (Fig. 6e), and (c) tumor size (T) and number of ELN combination test (Fig. 6f). The predicted survivability decreases with an increasing number of positive lymph nodes. However, collaborative staging (CS) tumor size or number of ELN does not substantially decrease the predicted survivability. The MLP model accurately predicts 93% of T-N cases, 19.6% of N-ENL cases, and 0% of T-ENL cases (Supplementary Table 18). In a 3-attribute varying test, the BCS-MLP model is evaluated using cases with T4 tumor size, N3 number of positive lymph nodes, and grade 4 condition at the same time (Fig. 6h). The BCS-MLP accurately predicts 90% of cases generated from surviving seeds (Class 1) and 98.9% of cases generated from death event seeds (Class 0).

Comparison of Wasserstein distances (WD)

We computed the Wasserstein distance between the original dataset and the generated test cases. Wasserstein distance captures the probability distribution shift given a metric space[22,23]. For IHM prediction, the Wasserstein distance between the original MIMIC-III training set and the synthesized multi-attribute tests is 33.4. This value is much larger than the Wasserstein distance (12.4) between the training data and test data split within the original MIMIC-III. In comparison, for BCS prediction, the distribution shift of the generated triple-attribute-based test cases from the original SEER dataset is smaller, with the Wasserstein distance being 9.8. The Wasserstein distance between the original SEER training set and the test set is 2.1 (Supplementary Table 19).

Impacts of resampling and reweighting methods

We trained and tested new ML models to assess the impact of resampling and reweighting methods on models' responsiveness. SMOTE and AdaSyn oversampling methods are used to enrich the minority prediction class. For MIMIC-III mortality prediction, resampled LSTM models are tested with our single-attribute critical zone test cases. Overall, the new models remain to have low responsiveness to high-risk patient conditions (Fig. 7a–c). Similar to the original models, models with resampling are still unable to recognize critical patient conditions. For example, LSTM with SMOTE consistently assigns low mortality risk scores to patients with critically high vitals (e.g., respiratory rate, temperature, systolic blood pressure). LSTM with AdaSyn is better at responding to elevated systolic blood pressures than the original model, however, it performs poorly in other tests. Tests with the eICU dataset give a similar or worse performance (Fig. 7d–f). For BCS and LCS prediction, new MLP models trained with SMOTE or AdaSyn oversampling methods exhibit similar trends as the original MLP model (Fig. 7g–l). The new models fail to recognize many critical cancerous conditions. In addition, for LCS prediction, SMOTE and AdaSyn methods make the LCS-MLP model less sensitive to increasing CS tumor size (Fig. 7j).

We also applied the reweighting approach to training. Supplementary Table 5 shows the cost parameters used. For mortality prediction, LSTM with reweighting gives comparable performance to the original model (Fig. 7a–f), except in one testing scenario. For eICU critically low systolic BP

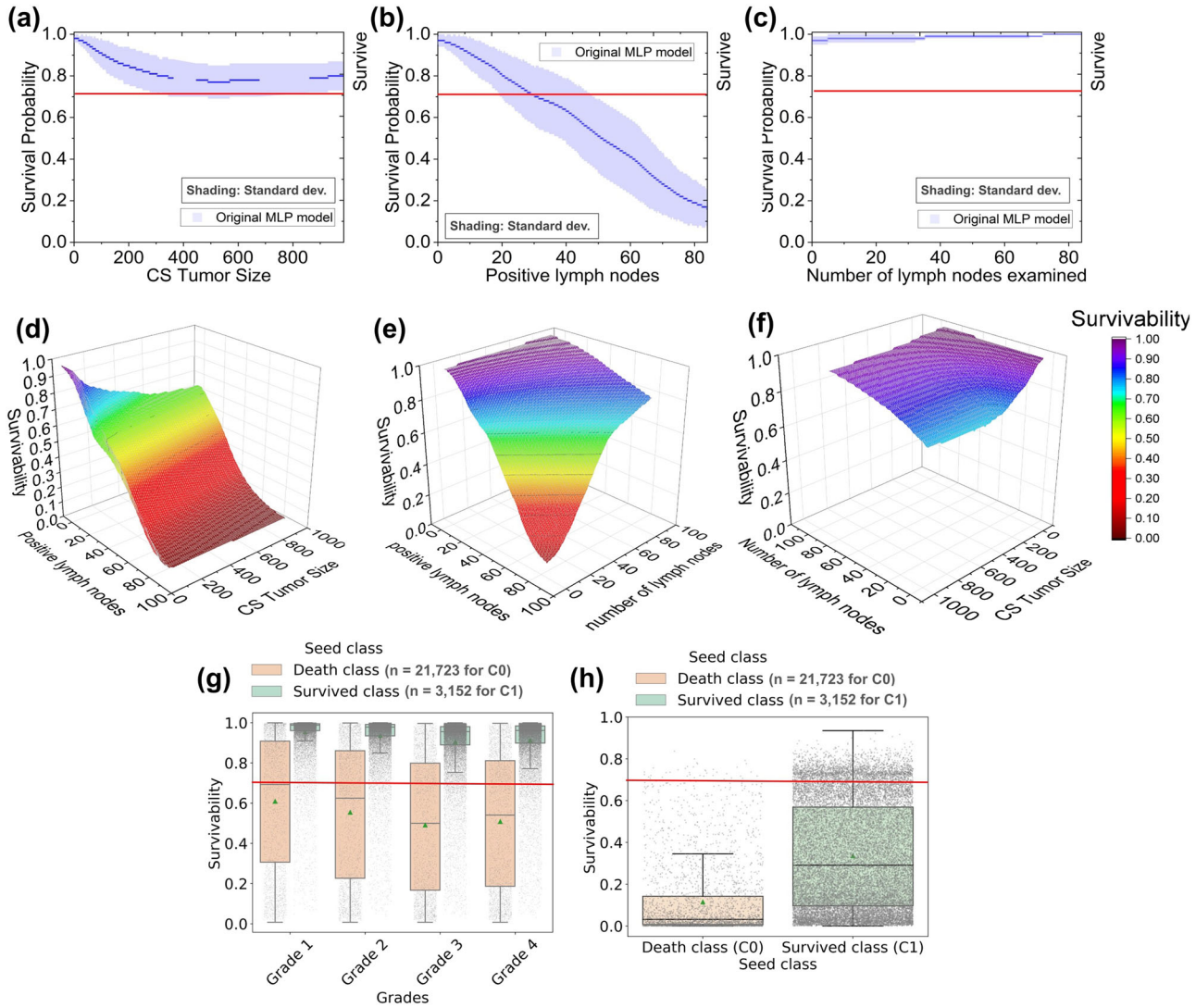

Fig. 6 | Predicted 5-year breast cancer survivability results of a multi-layer perceptron (MLP) model on test cases. Four major breast cancer screen attributes are involved, including CS tumor size, number of positive lymph nodes, number of lymph nodes examined, and grade. **a**–**c**, **g** Predicted survivability results on single-attribute varying test cases. The blue shaded area of **a**–**c** represents the standard deviation calculated from 3 independent experimental trials (models). **d**–**f** Predicted survivability results on double-attribute varying test cases. **h** Predicted survivability results on triple-attribute varying test cases involving CS tumor size, number of positive lymph nodes, and grade. In the boxplots (**g**) and (**h**) the horizontal line within the box represents the median value, while the box itself encompasses the interquartile range (IQR), containing the middle 50% of the data. The whiskers extend to the values within 1.5 times the IQR from the box (upper and lower quartiles). The green triangle point on the box represents the mean of the distribution. The points on boxplots (**g**) and (**h**) represent 21,723 from class 0 (C0) and 3152 samples from class 1 (C1).

tests, reweighted LSTM generates elevated risk scores and is slightly better at responding to abnormal patient conditions. However, reweighted LSTM performs worse for similar MIMIC-III test cases (Fig. 7c). For cancer survivability prediction, reweighting does not impact MLP's performance in most testing scenarios (Fig. 7g–l). For BCS test cases, the reweighted MLP model has slightly better responses to the increasing number of positive lymph nodes than the original MLP, however, it performs worse than the original MLP in terms of recognizing larger tumor sizes.

Responsiveness results of transformer models

The transformer models exhibit more responsiveness than LSTM in mortality prediction. They show elevated response in critically high zones for respiratory rate and systolic blood pressure, as well as in the critically low zone for temperature (Fig. 8). This trend is observed for the single-attribute test cases of both MIMIC-III and eICU datasets. In addition, transformer models recognize both critical zones of systolic blood pressure, yielding a desired U-shaped response curve (Fig. 8e, h). However, the transformer models fail to recognize critically low respiratory rates and critically high temperatures and exhibit low responsiveness to critically low systolic blood pressure. It also has delayed response to abnormally high respiratory rates. The transformer model's risk prediction fluctuates significantly for eICU test cases.

Discussion

Our findings highlight the importance of measuring how clinical ML models respond to serious patient conditions. Our results show that most ML models tested are unable to adequately respond to patients who are seriously ill, even when multiple vital signs are extremely abnormal. For time-sensitive IHM prediction, the lack of response to disease conditions is particularly troublesome. ML responsiveness is somewhat related to feature importance in some cases, e.g., the low responsiveness of LSTM to oxygen saturation tests (Fig. 3f and Supplementary Table 16) is consistent with that feature's low (15th) ranking (Supplementary Fig. S5). However, for high-ranking features such as glucose and temperature, ML responsiveness to

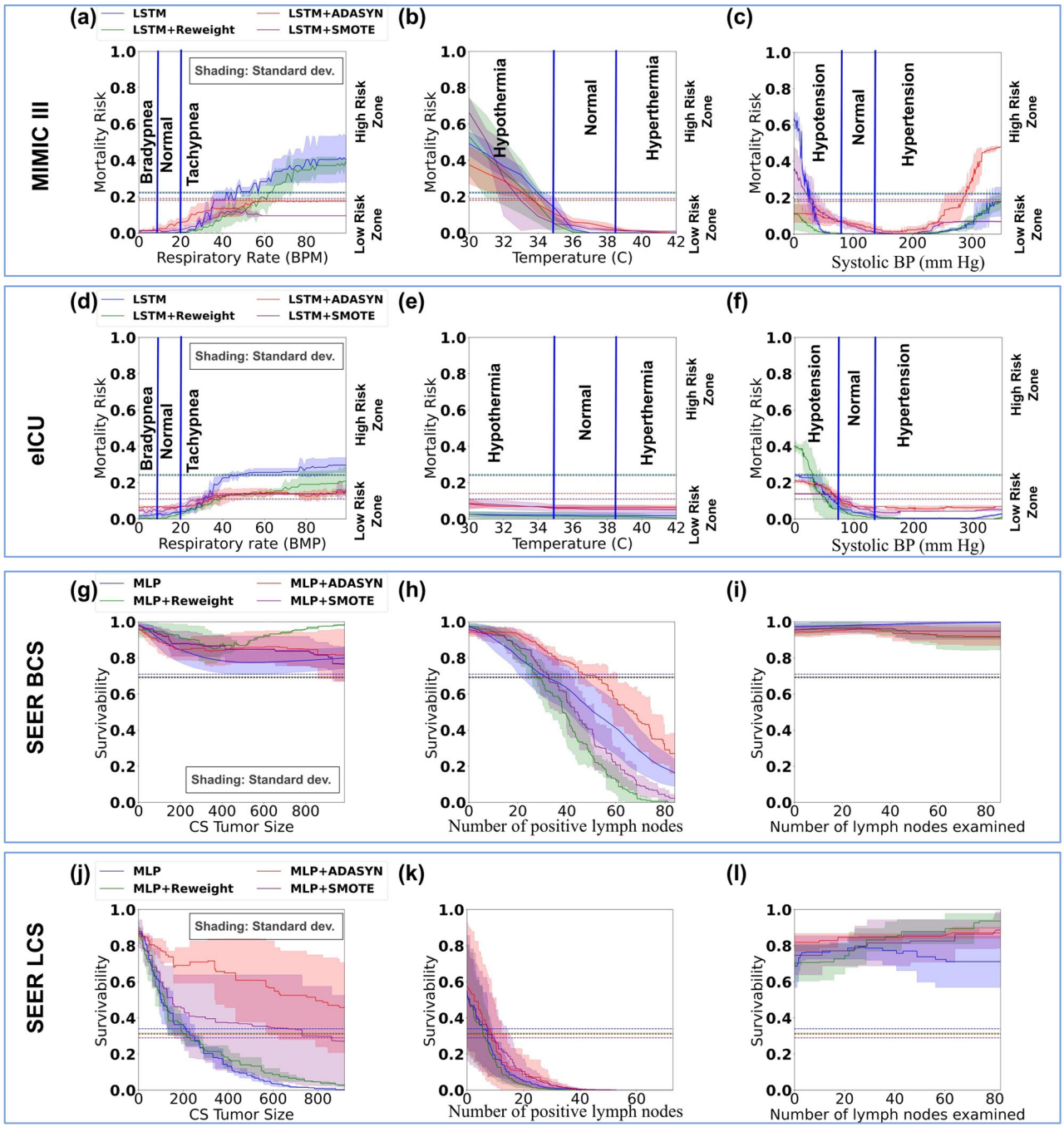

Fig. 7 | Performance comparison between the original machine learning models and the resampled (SMOTE or AdaSyn) or reweighted models under single-attribute varying tests. a–c and **d–f** Mortality risk prediction results by the original LSTM model and the resampled or reweighted LSTM models under MIMIC-III and eICU test cases for respiratory rate, temperature, and systolic blood pressure, respectively. **g–i** and **j–l** 5-year cancer survivability prediction results by the original MLP model and the resampled or reweighted MLP models under SEER BCS and LCS test cases for CS tumor size, the number of positive lymph nodes, and the number of lymph nodes examined, respectively. Horizontal dashed lines represent model-specific thresholds. The shading represents the standard deviation calculated from 3 independent experimental trials (models).

them is still inadequate. This poor responsiveness is also observed in the lack of responses in neural activation values (Fig. 3 and Supplementary Table 17) to important vital changes, such as extremely low respiratory rate or high body temperature.

New ML responsiveness metrics, especially for the healthcare domain, are urgently needed. ML responsiveness is a new problem. It differs from the well-studied ML robustness[37]. ML robustness aims to ensure model stability and the ability to resist sample perturbations so that small (maliciously injected) noises to samples cannot change the prediction results. Lipschitzness, a common ML robustness metric, measures the model's resilience to noisy data and perturbations[38]. However, for healthcare applications, optimizing Lipschitzness may lead to models being even more insensitive to changes in patient conditions, as adherence to Lipschitz continuity may hinder the model's ability to capture crucial input variations. In image and natural language domains, a common testing approach is adversarial attacks[39–42]. That testing approach involves intentionally manipulating input data to deceive the model's predictions and does not apply to our medical settings.

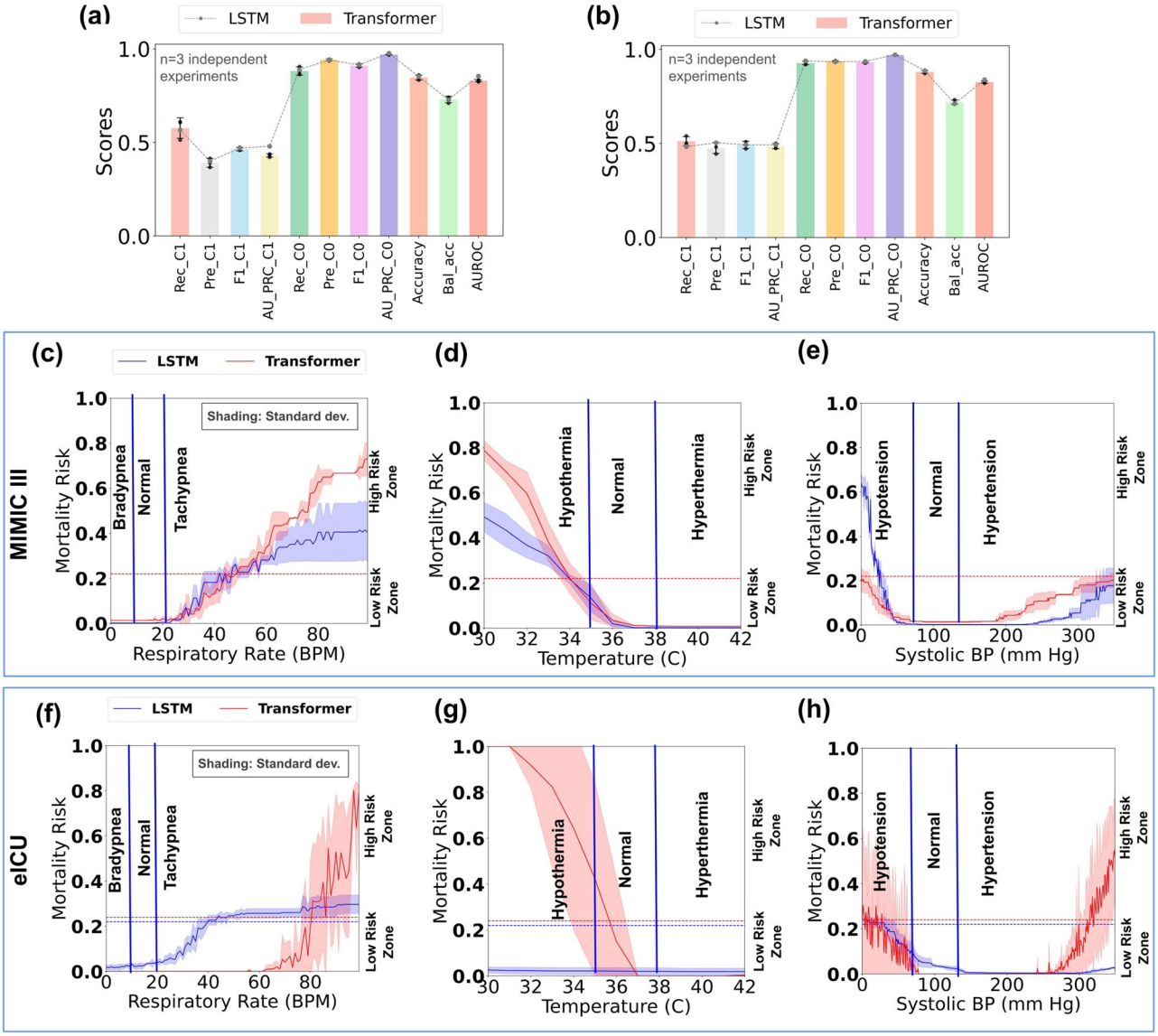

Fig. 8 | Performance and responsiveness of the transformer model compared with LSTM. Figures (**a**) and (**b**) show the various Class 1 (death) and Class 0 (survival) performance of the transformer model trained and tested on the original MIMIC-III and eICU datasets, respectively, with error bars indicating the standard deviation from three experimental trials. The dashed line represents the LSTM model's performance. Figure (**c**–**e**) shows the predicted mortality risk by the transformer model for respiratory rate, temperature, and systolic blood pressure on MIMIC-III single-attribute test cases, while (**f**–**h**) shows the same for eICU test cases. Horizontal dashed lines denote model-specific thresholds for mortality risk prediction. Rec_C1, Pre_C1, F1_C1, AU_PRC_C1, Rec_C0, Pre_C0, F1_C0, AU_PRC_C0, Accuracy, Bal_Acc, and AUROC stand for Recall Class 1, Precision Class 1, F1 score Class 1, Area Under the Precision-Recall Curve Class 1, Recall Class 0, Precision Class 0, F1 score Class 0, Area Under the Precision-Recall Curve Class 0, Accuracy, Balanced Accuracy, and Area under the Receiver Operating Curve, respectively. The shading represents the standard deviation calculated from 3 independent experimental trials (models).

Our results identified serious deficiencies in conventionally trained binary classification models in recognizing seriously abnormal medical conditions. For example, IHM prediction models fail to generate alerts for bradypnea (low respiratory rates) or hypoglycemia conditions (Fig. 3). Similarly, the models also consistently underestimate some of the mortality risks when given multiple abnormal vital time series in conjunction (Fig. 4). When given test cases representing various injury levels, neural network models (namely, LSTM and CW-LSTM) gave inconsistent risk predictions —assigning higher mortality risk (>0.5) to cases of moderate injury (e.g., GCS score 12), while assigning disproportionately lower risk (<0.05) to severe injuries (e.g., GCS score 7). The two neural network models exhibit insensitivity to changes in eye response (Fig. 2d, f). For most attributes, we found the training data's distribution is highly centered, not sufficiently representing high or low critical zones (Supplementary Fig. S1). Death and non-death cases exhibit somewhat similar value distributions, means, and standard deviations (Supplementary Table 14) for individual attributes, despite the drastically different outcome. ML methods produced by supervised training approaches are unable to recognize the meanings of vitals in dangerous zones. This semantic deficiency of ML models was also reported in image recognition studies, e.g., melanoma classification over-interpreting surgical skin markings[7]. We found similar kinds of semantic deficiencies in models predicting 5-year BCS. These findings indicate the importance of empirically assessing the trustworthiness of clinical ML models.

The conventional test set is limited in its distribution shift from the training data. For example, for IHM prediction, our generated multi-attribute test cases present a high Wasserstein distance (33.4) from the original MIMIC-III training data, much larger than the split test set's Wasserstein distance (12.4). A similar distribution shift pattern is observed for the BCS model. For triple-attribute test cases, the BCS

prediction model performs better (89–98% triple-attribute test accuracy) than IHM prediction models (6–69% multi-attribute test accuracy). This difference in accuracy may be partly due to the different distribution shifts in generated test data. There is a much smaller distribution shift in triple-attribute breast cancer test cases (Wasserstein distance 9.8) than in multi-vital test cases (Wasserstein distance 33.4), with respect to their original training data. The LSTM model is slightly better at recognizing multi-attribute test cases (45.7% accuracy) than single-attribute ones (37.8%) and CW-LSTM exhibits a similar pattern (69.3% vs. 22.4%, Supplementary Table 16). Multiple abnormal vitals likely provide more clues for the ML models to classify, whereas single isolated attribute changes appear more difficult. The poor performance of the ML models is somewhat expected because of the distribution shift between training data and our synthetic test data. Yet, these deficiencies are unacceptable from a clinical deployment perspective, as the test cases represent potential real-life medical conditions. Our work points out a fundamental limitation of pure data-driven machine-learning models, that is models purely trained by patient data do not perform well for tasks that require implicit medical knowledge (e.g., normal vital ranges).

For IHM prediction, all models have multiple deficiencies under single-attribute critical zone test cases and are unable to generate high enough risk predictions for serious patient conditions (Figs. 3, 8, and Supplementary Table 16). Out of all the dual critical zone attributes, only LSTM and transformer models exhibit U-shape risk curves for systolic blood pressure (Figs. 3e and 8e). The other risk curves are either monotonic or flat (Fig. 3). Transformer models implementing the parallel attention mechanism are known to be better at capturing the global context and dependencies in data than sequential models like LSTM. Indeed, transformers are more responsive to abnormal vitals than LSTM in single-attribute testing (Fig. 8). However, our results suggest advanced models alone are not sufficient, as there are still multiple unrecognized critical zones. For multiple attribute testing, LR performs the worst (Fig. 4 and Supplementary Table 16), mispredicting 93.7% of test cases. CW-LSTM's accuracy is the lowest (22.4%) in single-attribute testing, however, it gives the highest average accuracy (69.3%) for multiple-attribute testing. Brain injury-related GCS test cases (Fig. 2) involve simple categorical data (as opposed to numerical data). LR gives the best performance, generating appropriate and consistent risk estimates, and substantially outperforms the two neural network models (Supplementary Table 16). These results suggest that for categorical attributes such as GCS, a simpler model like LR may be more suitable than complex deep learning models, indicating the importance of evaluating a wide variety of ML models before clinical use. Deficiencies in ML responsiveness were also observed in the 5-year breast cancer prediction task—the MLP model gave an average of 48.9% prediction accuracy under our test case (Supplementary Table 18). This accuracy is much lower than the widely reported death class accuracy of 90% (standard deviation 0.45), which is based on the original test data from SEER[5,16].

The linear LR model tested is unsuitable for analyzing vitals due to multiple reasons. It reduces time series to statistical summaries (Supplementary Fig. S6c) and is unable to capture data dynamics. Linear LR is unable to model non-monotonic (e.g., U-shaped curve) features, as it responds monotonically to features. In multi-attribute tests, the model gives poor performance (6.2% accuracy, Supplementary Table 16), partly because of its many negative coefficients associated with attributes. 13 out of the 17 attributes have negative coefficients, e.g., the temperature is strongly inversely correlated with the predicted probability (Supplementary Fig. S6a), resulting in underestimated risk prediction. Gaussian Naive Bayes and KNN show weaker performance on the original test sets than the others (Supplementary Fig. S7) and, thus, are excluded from subsequent attribute tests. For completeness, model performance on the original test set was given in Supplementary Fig. S7.

For IHM prediction, the LSTM models are slightly better at recognizing deteriorating trends (average 44%, Fig. 5) than cases with steadily low or steadily high vitals in critical zones (average 36.5%, Figs. 3 and 4). When test cases contain 3 simultaneously deteriorating attributes, the models detected 56% of them on average, which is better than their performance of 41% on a single deteriorating attribute. When using LSTM gradient ascent to automatically generate multi-attribute deteriorating test cases, we found the resulting test cases all have significantly decreased oxygen saturation and body temperature values in the last 24 h. Because the gradient ascent process follows the shortest path within the loss function space of the model, these findings indicate that (i) oxygen saturation and body temperature are top LSTM features and (ii) the last 24 h (out of the entire 48-h timespan) are important in the model's decision-making process, which is also consistent with LR feature ranking (Supplementary Fig. S6b).

The BCS multilayer perceptron model (MLP) model exhibited responsiveness to critical attributes, such as tumor size and lymph node involvement (Fig. 6). For example, for N3 stage (extensive lymph node involvement) test cases, the model was able to raise alerts for 74.4% of them (Supplementary Table 18). The model also performed well (nearly 100% alerts) when all three critical features (T4 tumor size, N3 lymph node stage, and grade 4) were high. These observations suggest the MLP model's prediction capability in extreme cases is good. However, the model does not respond to severe tumor sizes (T3 stage), generating no alerts. The consistency in predicted survivability scores is also low, as the model generated slightly more alerts for grade 3 cancer (65.8%) than grade 4 terminal cancer (63.9%). This inconsistency may be the outcome of the imbalanced dataset (Supplementary Fig. S2), as the SEER dataset contains a total of 81,749 (death 15,628 and survived 66,121) grade 3 cases, while only 3002 (death 640 and survived 2362) grade 4 cases. The number of ELNs does not directly indicate a cancerous condition, thus, the model's lack of response to ELN is somewhat expected.

Despite overall good performance on the original test set (Supplementary Fig. S7), tree-based ensemble methods such as XGBoost, AdaBoost, and RF exhibit low responsiveness to critical zone tests (Supplementary Fig. S4). Ensemble methods perform much worse than MLP for SEER BCS and LCS settings, which is likely due to the sparsity in the one-hot encoded input space. The SEER dataset has much larger feature dimensions (56 for BCS and 47 for LCS) than the MIMIC III and eICU time-series data (17 features). Using the one-hot encoding to encode categorical features leads to an expansive number of sparse encoded representations (1423 for encoded BCS and 1315 for encoded LCS), posing challenges to tree-based models.

One clinical mitigation is to deploy a filter-then-predict workflow where domain-specific rules are first applied to identify cases with obvious disease conditions. Thus, corner-case scenarios will never reach ML models. However, designing such rule-based classifiers, especially under time-series data, is challenging and may require substantial manual efforts. A more efficient approach is out-of-distribution detection, which identifies cases that present a large distribution shift from the model's training data. Existing solutions for detecting out-of-distribution images[43] cannot be directly applied to clinical settings. For medical applications, out-of-distribution patient cases still need to be examined and handled. An overly strict detection may produce too many such out-of-distribution cases for downstream examination. Finding the right balance will facilitate clinical translation.

A promising direction is medical foundation models based on clinical large language models (LLMs)[44–47]. Our findings suggest that statistical machine-learning models solely trained from patient data are grossly inadequate. They are unable to capture basic clinical knowledge, e.g., patients with extremely low GCS values have a high mortality risk. LLMs are likely able to recognize common sense health conditions and serve as a filter mechanism before ML classifiers. However, it is crucial to quantitatively characterize the trustworthiness of medical LLMs before clinical adoption. Our work suggests the urgent need for innovative clinical decision-making workflows, as existing models solely trained from patient samples are extremely limited. For interpreting ML results, a human-friendly interface is also important. Conventional interpretability techniques, such as SHapley Additive exPlanations (SHAP)[48], Local Interpretable Model-Agnostic Explanations[49], or TRUSTEE[50], were designed for ML experts, not for clinicians. Therefore, these tools cannot be directly used in clinical settings.

An innovative clinical workflow needs to place generative AI as the final component to generate narrative explanations based on ML predictions and interpretability results. An interesting research direction is how to fine-tune LLMs for these specific tasks.

The boost provided by conventional resampling and reweighting methods is very limited (Fig. 7). Under some scenarios (e.g., tachypnea and increasing CS tumor size), they may perform even worse than the original models. This poor performance is expected, as these methods rely on existing minority class samples in the training set, which are limited in their ranges and variations. The root problem is that the space of all possible minority class samples is vast. Attempting to cover all or most of them through training data engineering (such as oversampling) is infeasible. Thus, data engineering does not appear to be a feasible direction for the ML responsiveness problem. A more promising approach is to directly encode medical semantics into the clinical decision workflow as discussed above.

Our work provides the first look into ML responsiveness. Comprehensive measurement studies in other medical settings are needed. Our gradient ascent testing methodology can be extended to other health conditions (e.g., rare diseases or comorbidities). Scalability is the key to testing in medicine, because of the complex high-dimensional space. Innovative methods that prioritize testing are needed to reveal the most critical blind spots in a model.

Data availability

The MIMIC-III, eICU, and SEER datasets used in this study are existing datasets available to researchers. They can be requested at their original sites after completing proper training. Parties interested in data access should visit the MIMIC-III website (https://mimic.physionet.org/gettingstarted/access/), eICU website (https://eicu-crd.mit.edu/) and the SEER website (https://seer.cancer.gov/data/access.html) to submit access requests. Because our test cases are generated from these access-controlled datasets, they cannot be publicly released. However, we have released the code for reproducing all our test cases. All the source data for the main figures (as a Microsoft Excel file) is available as Supplementary Data 1.

Code availability

We have released all our code publicly on GitHub, which can be used to generate the test cases and reproduce our experiments. https://github.com/PiasTanmoy/TRUSTWORTHY-ML (https://doi.org/10.5281/zenodo.14254248)[51].

Received: 11 June 2024; Accepted: 17 February 2025;
Published online: 11 March 2025

References

1. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39 (2018).
2. Sennaar, K. How America's 5 top hospitals are using machine learning today. *Emerj* https://emerj.com/ai-sector-overviews/top-5-hospitals-using-machine-learning (2020).
3. Sendak, M. P. et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med. Inform.* **8**, e15182 (2020).
4. Zaribafzadeh, H. et al. Development, deployment, and implementation of a machine learning surgical case length prediction model and prospective evaluation. *Ann. Surg.* **278**, 890–895 (2023).
5. Afrose, S., Song, W., Nemeroff, C. B., Lu, C. & Yao, D. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Commun. Med.* **2**, 111 (2022).
6. Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).
7. Winkler, J. K. et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**, 1135–1141 (2019).
8. Liang, W. et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **4**, 669–677 (2022).
9. Tian, Y., Pei, K., Jana, S. & Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. *Proceedings of the 40th International Conference on Software Engineering* 303–314 (Association for Computing Machinery, 2018).
10. Pei, K., Cao, Y., Yang, J. & Jana, S. Deepxplore: automated whitebox testing of deep learning systems. *Proceedings of the 26th Symposium on Operating Systems Principles* 1–18 (Association for Computing Machinery, 2017).
11. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. data* **6**, 96 (2019).
12. Johnson, A., Pollard, T. & Mark, R. MIMIC-III clinical database (version 1.4). *PhysioNet* **10**, 2 (2016).
13. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
14. Sheikhalishahi, S., Balaraman, V. & Osmani, V. Benchmarking machine learning models on multi-centre eICU critical care dataset. *PloS ONE* **15**, e0235424 (2020).
15. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 1–13 (2018).
16. Hegselmann, S., Gruelich, L., Varghese, J. & Dugas, M. Reproducible survival prediction with SEER cancer data. *Proceedings of the Machine Learning for Healthcare Conference*, vol. 85, 49–66 (PMLR, 2018).
17. Khadanga, S., Aggarwal, K., Joty, S. & Srivastava, J. Using clinical notes with time series data for ICU management. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 6432–6437 (Association for Computational Linguistics, 2019).
18. Deznabi, I., Iyyer, M. & Fiterau, M. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* 4026–4031 (Association for Computational Linguistics, 2021).
19. Zhang, H., Singh, H., Ghassemi, M. & Joshi, S. Why did the model fail?": attributing model performance changes to distribution shifts. *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, 41550–41578 (PMLR, 2023).
20. Zhou, H., Chen, Y. & Lipton, Z. Evaluating model performance in medical datasets over time. *Proceedings of the Conference on Health, Inference, and Learning*, vol. 209, 498–508 (PMLR, 2023).
21. Mienye, I. D. & Sun, Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Inform. Med. Unlocked* **25**, 100690 (2021).
22. Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D. & Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal Process. Mag.* **34**, 43–59 (2017).
23. Villani, C. *Topics in Optimal Transportation* Vol. 58 (American Mathematical Soc., 2021).
24. Jain, S. & Iverson, L. M. Glasgow coma scale. (2018).
25. *Stages of breast cancer: Understand breast cancer staging — American Cancer Society*, accessed 31 October 2024 <https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html>.
26. Shanmugam, C. et al. Evaluation of lymph node numbers for adequate staging of Stage II and III colon cancer. *J. Hematol. Oncol.* **4**, 1–9 (2011).

27. Yong, J., Ding, B., Dong, Y. & Yang, M. Impact of examined lymph node number on lymph node status and prognosis in FIGO stage IB-IIA cervical squamous cell carcinoma: a population-based study. *Front. Oncol.* **12**, 994105 (2022).
28. Choi, H. K., Law, W. L. & Poon, J. T. The optimal number of lymph nodes examined in stage II colorectal cancer and its impact of on outcomes. *BMC Cancer* **10**, 1–7 (2010).
29. Wu, Q. et al. Impact of inadequate number of lymph nodes examined on survival in stage II colon cancer. *Front. Oncol.* **11**, 736678 (2021).
30. Sun, L., Li, P., Ren, H., Liu, G. & Sun, L. Quantifying the number of lymph nodes for examination in breast cancer. *J. Int. Med. Res.* **48**, 0300060519879594 (2020).
31. Chi, H., Zhang, C., Wang, H. & Wang, Z. The appropriate number of ELNs for lymph node negative breast cancer patients underwent MRM: a population-based study. *Oncotarget* **8**, 65668 (2017).
32. *Understanding blood pressure readings—American Heart Association*, accessed 31 October 2024 <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>.
33. *Vital signs (body temperature, pulse rate, respiration rate, blood pressure) —Johns Hopkins Medicine*, accessed 30 October 2024 <https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure>.
34. *Vital signs (body temperature, pulse rate, respiration rate, blood pressure)—University of Rochester Medical Center,* accessed 31 October 2024. https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=85&ContentID=P00866.
35. *Vital Signs - Cleveland Clinic*, accessed 30 October 2024 <https://my.clevelandclinic.org/health/articles/10881-vital-signs>.
36. SapraA., Malik, A. & Bhandari, P. *Vital SIGN Assessment*. *In*: *StatPearls* (Treasure Island, 2023).
37. Mirman, M., Gehr, T. & Vechev, M. Differentiable abstract interpretation for provably robust neural networks. *Proceedings of the International Conference on Machine Learning*, vol. 80, 3578–3586 (PMLR, 2018).
38. Qin, Y. et al. Stolen Risks of Models with Security Properties. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* 756–770 (Association for Computing Machinery, 2023).
39. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. & Tygar, J. D. Adversarial machine learning. *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* 43–58 (Association for Computing Machinery, 2011).
40. Tygar, J. Adversarial machine learning. *IEEE Internet Comput.* **15**, 4–6 (2011).
41. Finlayson, S. G., Chung, H. W., Kohane, I. S. & Beam, A. L. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296* (2018).
42. Newaz, A. I., Haque, N. I., Sikder, A. K., Rahman, M. A. & Uluagac, A. S. in *GLOBECOM 2020–2020 IEEE Global Communications Conference* 1–6 (IEEE, 2020).
43. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: a survey. *Int. J. Comput. Vis.* **132**, 5635–5662 (2024).
44. Ong, J. C. L. et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. *Cell Rep. Med.* **5**, 101356 (2024).
45. Yang, X. et al. A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194 (2022).
46. Peng, C. et al. A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **6**, 210 (2023).
47. Editorial. How to support the transition to AI-powered healthcare. *Nat. Med.* **30**, 609–610 (2024).
48. Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. *Proceedings of the Advances in Neural Information Processing Systems 30* (Curran Associates, Inc. 2017).
49. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, 2016).
50. Jacobs, A. S. et al. AI/ML for network security: The emperor has no clothes. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* 1537–1551 (Association for Computing Machinery, 2022).
51. Pias, T. S. Code: low responsiveness of machine learning models to critical or deteriorating health conditions. *Zendo*. https://doi.org/10.5281/zenodo.14254248 (2024).

Author contributions

D.Y. conceived and designed the study. T.P. and D.Y. conceived the gradient-based test generation method. D.Y. and T.P. conceived the new metrics introduced. T.P. conducted the experiments and analyzed the data, including selecting seeds for test case generation and interviewing medical experts. S.A. performed the initial neuron heat map study. X.D. guided T.P. to perform the WD statistical analysis. M.T. and I.T. provided medical feedback on the mortality risks of generated test cases through Zoom interviews. D.Y. and T.P. wrote the manuscript. C.B.N. provided strategic guidance. All authors proofread the manuscript and provided feedback.

Competing interests

C.B.N. is supported by the National Institutes of Health, the National Institute of Mental Health, and the National Institute of Alcohol Abuse and Alcoholism. Charles Nemeroff is a consultant for ANeuroTech (division Anima BV), Janssen Research and Development, BioXcel Therapeutics, Engrail Therapeutics, Clexio Biosciences LTD, EmbarkNeuro, Galen Mental Health LLC, Goodcap Pharmaceuticals, ITI Inc, LUCY Scientific Discovery, Relmada Therapeutics, Sage Therapeutics, Senseye Inc, Precisement Health, Autobahn Therapeutics Inc, EMA Wellness, Skyland Trails, Denovo Biopharma, Alvogen, Acadia Pharmaceuticals, Inc., and the Brain & Behavior Research Foundation; owns the following patents: Method and devices for transdermal delivery of lithium (US 6,375,990B1), Method of assessing antidepressant drug therapy via transport inhibition of monoamine neurotransmitters by ex vivo assay (US 7,148,027B2), Compounds, Compositions, Methods of Synthesis, and Methods of Treatment (CRF Receptor Binding Ligand) (US 8,551, 996 B2); owns stock in Corcept Therapeutics Company, EMA Wellness, Precisement Health, Relmada Therapeutics, Signant Health, Galen Mental Health LLC, and Senseye Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00775-0.

Correspondence and requests for materials should be addressed to Danfeng Daphne Yao.

Peer review information *Communications Medicine* thanks Allan Tucker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.